The Many Definitions of Test Bias

RONALD L. FLAUGHER *Educational Testing Service*
Princeton, New Jersey

ABSTRACT: *The definition of test bias—the inventory of the ways in which the term is used—has many, widely disparate aspects frequently stemming from entirely different universes of discourse. This article attempts a review of the status of each of these. It seems essential to keep all of these various aspects in mind, for we continually run the risk of losing perspective on our research when we settle on one operational definition of test bias and then proceed to forget that it is only that. No matter what definition we use, because the concept is a public one we are never going to encompass all that it contains. The research on these various aspects is some of the more exciting and significant being done in psychology today, but let us not be confused about whether one of them is the real issue. As it happens, they are* all *the real issue.*

A major theme of this article is that the definition of test bias—the inventory of the ways in which the term is used—has many aspects. These aspects are widely disparate and frequently stem from entirely different universes of discourse.

People still demand an answer to the question "Are tests biased?" and expect the answer to be either "yes" or "no" without going to the trouble of defining just what that question can mean. When, instead, the questioner finds that he or she has touched off a discussion of the sort that follows in this article, it is often interpreted as evasiveness. Nevertheless, an accurate answer to this question must be an involved one.

It might be argued that we could simplify things by discarding the popular conceptions of the problem, that we as researchers cannot be held responsible for the misconceptions and fallible interpretations of the less sophisticated. While such might well be the case in other scientific pursuits, it happens that this issue of bias in testing is currently appearing in public forums, including courts of law, and decisions are being made that have an impact on critical issues such as who shall be educated and who shall be employed. It does not appear, in other words, that we have the luxury of limiting the definition of test bias to what we would like it to be. The public is concerned, and it behooves us to become concerned with the public. Their definitions must be our definitions.

I would like to elaborate on the complexity of the issue of test bias in the remainder of this article and to set our past research endeavors and, more important, those of the future in perspective. Some aspects of the phenomenon are quite accessible to our research techniques and others simply are not; some, though accessible, turn out to encompass but a very tiny part of the total phenomenon. The fact that they are accessible, however, has a magnifying effect on their apparent significance.

This task of reviewing the various aspects and interpretations of test bias, then, can make two significant contributions: (1) It can help us order our own research priorities so that progress is attempted where progress is possible, and (2) it can enable us to communicate meaningfully when we are asked to participate in public deliberations on the topic.

Very briefly, the points I make are these: First, an important complicating aspect is the widespread failure to interpret test scores appropriately, most notably in making the distinction between aptitude and achievement tests. Second, some research is yielding disappointing results in spite of its initial promise. Third, other research is leading us to the examination of our own value systems, an unfamiliar task for most of us. Fourth, still other research may apply to the concept of test bias in ways we didn't expect. I will make these points by considering each of a number of aspects of the concept of test bias in turn.

Test Bias in Achievement Tests

One interesting aspect of the issue of test bias is its existence or nonexistence in achievement tests. One would have thought this issue would have

---

Requests for reprints should be sent to Ronald L. Flaugher, Center for Occupational and Professional Assessment, Educational Testing Service, Princeton, New Jersey 08540.

0003-066X/78/3307-0671$00.75

been put to rest by some earlier papers, such as that of Ebel (Note 1), but instead the debate rages on. I believe that the indestructible nature of the debate has its origins in the emotionality attached to test-score interpretation—that is, to the use to which the test scores are put. And part of the problem lies in the failure to make the classic distinction between the aptitude test and the achievement test. The two do in fact frequently overlap in content, of course, because it is simply not the case that current aptitude can be measured in isolation from past achievement; the correlations of the two so-named tests are frequently positive and high, but correlation is not causation, and that old precaution may need to be revived again here. But there is a great deal of difference, a difference of great social significance, between the two types of test in the kind of legitimate interpretations that are put on the results.

If a test is an achievement test, and is *interpreted* as an achievement test, then a high score means that a course of instruction has been successfully conveyed and that further educational effort in that area is unnecessary. On the other hand, a low score on that same test means that additional educational effort, perhaps more of the same or of some other kind, is called for, since the achievement has not occurred.

But what if the test is seen as a test of pure aptitude? In that case, instead of measuring accomplishment, the intent is to measure the *capacity* for accomplishment (Nunnally, 1959, p. 265). Thus, a high score bodes very well for the test taker, but a low score may be interpreted as an indication that there is insufficient capacity on that test taker's part to achieve; therefore, any additional educational effort would be wasted.

So, in the distinction between achievement and aptitude tests, or more correctly, in the distinction between the *interpretation* of achievement and aptitude tests, lies a social decision of great significance. If test performance is low and it is seen as an index of achievement, then there is pressure to increase the application of society's educational resources to improve that achievement. If the test result is seen as an index of aptitude, however, then the same low test score may be seen as a justification of the *withdrawal* of educational resources. In the one case, society assumes greater responsibility; in the other, it escapes or denies responsibility.

I have seen the failure to make this distinction lead to great emotionality and failure to face serious social problems. Low achievement scores reported for groups of minority students have led to demands not for the improvement of the educational system but for the abandonment of those "lying" tests, which are seen as indicating that the capacity to achieve does not exist.

Achievement tests are certainly capable of being poorly composed, but when the many carefully constructed and widely used tests are actually inspected for content, almost always that content is found to consist of legitimate samplings of quite uncontroversial educational goals, such as the ability to read a popular newspaper or to calculate simple mathematical problems. But the emotionality stemming from the confusion over the proper interpretation of the test scores sometimes leads minority spokespersons to take the position that those tests are asking the *wrong questions*. This implies that somewhere there are some *right* questions and that in fact there is satisfactory achievement taking place on the part of our minority students. Since the tests are asking the wrong questions, minority spokespersons argue that achievement is happening but not being documented and that the tests are misleading everyone. It is as if they are saying that our inner-city schools are in fact being quite successful after all.

But the interesting thing is that some of the same persons who will at least implicitly stand by this argument at some times will at other times be vocal critics of that same school system, claiming that it has failed the minority population. The strain on the logic of the argument occurs when those same low test scores are used to prove the point.

I submit that this particular sort of contradictory reasoning must be dealt with before appropriate social actions can be decided upon. And, closer to home, such issues must be clarified before we can decide upon the proper directions for our research and development efforts—do we go out in pursuit of those "right questions," the undocumented achievement that some claim is occurring in those schools, or not?

There are many powerful analogies that have been invoked with the intention of countering the arguments for a moratorium on testing: You don't solve the problem by assaulting the messenger who brings the bad news; you don't destroy the thermometer that tells you that a fever exists; you don't cure malnutrition by throwing out the scales that identify the underweight babies. Yet the analogies don't seem to persuade the critics, and the demands continue. The reason appears to be that this process is seen not as a diagnosis that will lead

toward attempts to cure the illness, but rather, because of the aptitude–achievement confusion, as an official certification that no help is possible. If the messenger is seen as working for the enemy, perhaps hostility is appropriate. If the high fever or presence of malnutrition is diagnosed as a terminal case, perhaps other opinions should be sought. We need to see just why there is this anger and emotionality before we can clarify the bias issue.

Test Bias as Mean Differences

It is tempting to exclude this particular definition of test bias from the discussion, since it is so easy to point out that mean differences, in themselves, are simply not a legitimate standard for identifying bias. But it must be included because many spokespersons of the minority cause start with that premise. This can be better understood if we view it as the merging of two facets of a problem: (1) a desirable goal, with (2) the evidence of that desirable goal.

Knowing what we do about the relative status, socioeconomic and otherwise, of ethnic minorities in the United States, it would be surprising if most kinds of tests didn't show mean differences in favor of the majority group. It would have to be a peculiar kind of test indeed to fail to reflect the disparities and differing advantages that are so evident through other sources. Yet many critics of testing merge the concept of equality of *opportunity*, which is certainly a legitimate goal to be sought, with the concept of equality of *result;* but it is only results that the tests in fact measure. The existing discrepancy is evidence that the legitimate goal has not been attained; to accept the discrepancy instead as evidence of test bias is to deflect attention from the pursuit of that legitimate goal.

Once again, this may be an example of the failure to make the distinction between achievement, which is a reflection of past accomplishment, with aptitude, defined as a statement about capacity to achieve in the future. Certainly, denial of capacity in an entire ethnic group would raise the question of bias, but that is not what mean differences in our tests are saying—or even can say. The myth of the achievement-free aptitude test will probably always be with us, but it is only a myth and serves to complicate our communications enormously.

Test Bias as Overinterpretation

One seemingly advantageous characteristic that tests have is the capacity to instill in interpreters the belief that they are assessing far more than they really are and, what is often something different, far more than what they claim to be assessing. One approach to understanding this problem is to view all the worthwhile attributes of human behavior as comprising a wide spectrum, and then to consider just what range within that spectrum can be assessed by the administration of our objective tests. It has become clear that a great discrepancy exists between how much and what part of that spectrum the public thinks we can measure and what parts we can in fact lay claim to measuring. The error on the public's part is in the direction of believing that a much greater portion of the spectrum is covered than is in fact covered. We do measure certain aspects of the spectrum quite well, generally those related to academic matters, but these comprise but a very narrow band within a spectrum that includes all worthwhile traits that make up human behavior. Adding to the problem is the fact that we do such a good job on that narrow portion; we are so accurate in measuring that portion that it is tempting to exaggerate its importance—since we can't measure the other important aspects of humanity, then we concentrate on those things we can measure. Since we can't measure all of the important things, we consider what we can measure all-important.

This tendency to overinterpretation, unfortunately, is not totally ascribable to wishful thinking on the part of the uninformed public. Wishful thinking on the part of testers and researchers also causes a stretching of the interpretation across very questionable areas of the spectrum. Ebel (Note 2) has pointed out that tests are frequently named not for the kind of tasks they present but for what they were intended to measure. A test of commonly encountered problems will not be called that but, rather, a test of "practical judgment." If the examinee is asked to offer suggestions, the word *creativity* will probably appear in the title. And a test of verbal, numerical, and symbolic problems is far less likely to be called a "problems" test than a test of intelligence. The overinterpretation that occurs is not without its encouragement within the profession. But when these tests are represented as assessments of highly valued parts of the spectrum, the issue of test bias is legitimately raised, since it is a great leap from being unable to work a few problems on a pencil-and-paper test to being declared lacking in practical judgment.

Test Bias as Sexism

In most respects, the question of fairness to women can be treated identically with that for ethnic minorities. One unique characteristic, however, serves to add still another dimension to the complexity of a complete definition of test bias. More than any other groups' concerns, the women's quarrel involves the very language itself, in addition to questions of the fairness of selection models, differential validity, and the like. "All of our language has a distinct masculine bias, including, in particular, the generic use of male nouns and pronouns when the content refers to both sexes (American Psychological Association, 1977). Tests are just another one of the users of the language and supposedly are no more prone to this sort of bias than any of the other users. It may be considered an indication of the importance that tests are seen to have in our society that so much attention has been given to the language usage in them—along with school textbooks—to assure that the cultural components reflected in them do not perpetuate the stereotypes of the past. Tests may not be any more guilty of such bias than other users of the language, but they are seen as an appropriate medium through which a desirable social change can be effected, a change not confined to the function of tests themselves.

Test Bias as Single-Group or Differential Validity

These issues are the operational definitions of the question of whether the same test can predict equally well for two identifiable groups—in this case, ethnic groups within the United States. There have been numerous studies, and declarations by many investigators, that single-group[1] and differential validity do not exist (Boehm, 1972; Campbell, Crooks, Mahoney, & Rock, 1973; Cleary, Humphreys, Kendrick, & Wesman, 1975; Flaugher, 1974; Humphreys, 1973; O'Conner, Wexley, & Alexander, 1975; Schmidt, Berner, & Hunter, 1973; Stanley, 1971). In two cases, these declarations represented reversals of earlier beliefs (Guion, 1972; Wallace, 1972). Nevertheless, a very recent edition of the *Journal of Applied Psychology* devoted three articles to this same topic, one concluding that ethnic differences in validity are not a "pseudoproblem" (Katzell & Dyer, 1977), one concluding that more and better research is needed before deciding (Bartlett, Bobko, & Pine, 1977), and one concluding that the reported positive findings can be regarded as methodological artifacts and that differential validity is nonexistent (Boehm, 1977).

The most sensible position to take on this matter would seem to be that even if single-group and differential validity "exist"—in the peculiar sense of that word as it is applied here—then the fact that they are so elusive, difficult to detect, and debatable is good evidence that they are not very potent phenomena relative to all the other possible sources of problems in the interaction of minorities and testing. Rather than continue this debate, it seems more appropriate to turn our energies to aspects that have a greater impact on real-life decisions.

Test Bias as Content

One of the first things people think of regarding test bias is the content of the test, the items that make it up. A biased test is one that contains questions that in some sense are "unfair" to some subgroup of the population. It appears that a great deal of the discussion on this aspect of test bias stems from confusion about the function of the test, the reason it is being taken, and the interpretation being put on the results. I have discussed these issues with groups that have used statements such as "Of course those tests are biased—I took one the other day, and it asked for definitions of words I had never even seen before." Another version of this statement is, "The tests are biased because they are asking questions that some kids have not even had the opportunity to learn."

The rejoinder to this kind of reasoning is apparent to us—the test is being given to see how many of these particular items can be answered, and regardless of the reason for the inability to answer them, whether through restricted opportunity to learn certain content or failure to utilize that opportunity, it is, for the moment, only the information about the ability or inability to answer these questions that is being sought, not the reason why.

Again, it is advisable here to reiterate the difference between aptitude and achievement testing

[1] The term *single-group validity* has come to refer, in this setting, to the finding of a validity coefficient significantly different from zero for one ethnic group but not the other; differential validity refers to the finding of a significant difference *between* two validity coefficients.

and to clarify the position of the persons objecting to the content of a given test. The public interpretation very frequently seems to be that these tests should be opportunities for the test takers to reveal whatever knowledge they possess, regardless of its nature or form. A fair test is one which asks all the questions that the taker can answer; a biased test is one that fails to ask a question the taker could have answered or that fails to reveal all of the strengths possessed. So the expression might be heard, "The test was biased because it did not ask what I know and did ask things I did not know."

This general definition of bias as particular content has led to elaborate procedures aimed at "debiasing" existing tests by having them scrutinized by a panel of experts chosen for their ability to recognize test items that are unfair to particular subgroups and then eliminating the items so identified from the test. Quite a number of attempts have been made, but the results have been documented only rarely. However, the results reported by Bianchini (Note 3) seem to be typical. In brief, the results indicated that eliminating from the test the items judged to be biased—in this case 13 of 82 items on a widely used elementary school reading test—did not improve the performance of schools with high minority populations relative to their performance on the original "biased" version. Further, informal experience with other such attempts has given another indication of the futility of this approach, in that the degree of agreement among the members of the panel of experts tends to be very low. When the issue is whether or not the child is able to read, it seems that this issue of asking the wrong questions just does not apply, at least in tests that have been constructed with some care in the first place.

There is another method of defining a biased item, one using a strictly empirical approach. By obtaining ethnic identity from the test taker, an analysis can be performed that compares the item-performance statistics group by group. If a particular item is extraordinarily difficult for the minority-group members relative to the difficulty of other items in the same test, then that item is a good candidate for suspicion of this kind of bias. The hope is that the particular content of a group of such items can be identified and that this will permit the avoidance of such content in the future. But in fact very little content similarity among such items has been found. A number of such studies have been conducted with generally disappointing results (Flaugher, 1974). This raises the question of whether the idiosyncratic performance on the item is simply a chance phenomenon; the next step is to pursue opportunities to examine the same item in a second, similar test administration to see if performance on it is similar in the new context.

But it is not absolutely necessary to question the content of the item; one tactic is simply to eliminate the idiosyncratic items from the test and to rescore. Schrader and I (Flaugher & Schrader, 1978) did just that with a form of the Scholastic Aptitude Test, and once again the results were discouraging. The test was rescored on the same group, dropping 12 items from the Verbal section and 15 from the Math section. This succeeded mostly in making the test considerably more difficult for everyone, since many of the items that showed the widest discrepancy between groups were moderate to low in overall difficulty. Further, the size of the discrepancy between the mean scores for the two groups was affected only slightly, suggesting that even the adjustment of the test content on this blind, empirical basis was not sufficient to make any important difference.

So the research picture is not promising in the area of test bias as content. Of course it would be possible to construct a test so carelessly and with so little sensitivity that some items would in fact be unfair to some subgroups, and numerous investigators have pinpointed individual instances of such things occurring. Considering the importance of testing, such scrutinizing of popular tests should certainly be continued. But the amount of change in test scores and the resulting decision changes that appear to be possible through this approach must be recognized as rather small, certainly small in comparison to the great amount of denunciation and criticism that testing is undergoing over the issue of bias. Something more is involved beyond simply asking the wrong questions.

Test Bias as the Selection Model

One dramatic new development is that of models of fairness in selection, which was not even considered a debatable issue less than a decade ago. Since that time, details of the various models have been worked over very thoroughly (see, for example, Petersen & Novick, 1976), and I will not repeat them here. There are issues of perspective here as well, however. For one thing, although the various papers discussing these selection models are often

grouped under the topic of "test" bias, the more accurate term, of course, would be *selection bias*—the fact that the predictor variable being employed is a pencil-and-paper test is only incidental to the characteristics of the various models. The same considerations would apply, and the debates rage, if any other kind of predictor variable were to be used in place of the test, such as essays, structured interviews, past academic records, or whatever. Actually, because of the reduced reliability characterizing virtually all of those alternative measures, there would in fact be an impact on those decision models, in the form of making the discrepancies among them more extreme. But it should be kept in mind that those models of selection have to do with issues other than the predictor used—issues that are in themselves also quite legitimate aspects of the total concept of test bias.

One major advance in our thinking that can be attributed to the development of these various models is that they are now seen to be mathematical expressions of particular value systems, value systems which themselves, potentially at least, can be arrived at by any means whatsoever through the application of whatever criteria are felt to be appropriate whether they be mathematical, ethical, or social. The search for the *really* fair model will therefore be a search for a set of values upon which everyone can agree. The prospects for that are dim, of course. The point is that there can be no single standard for selection fairness that has been determined objectively, only systems that are more or less popular with decision makers. And those decision makers have the task of trying the available models, not with the conviction that some one of them is "true," but in order to determine whether that model does for them what they have decided, on other bases, they want done in the selection process. The results of such actual applications can be interesting, as in, for instance, the study by Breland and Ironson (1976) which found that one of the models that was considered quite favorable to minorities was still less so than the one that was arbitrarily imposed (that is, without the benefits of our psychometric manipulations and deliberations) by the University of Washington in the famous DeFunis case. This suggests that quite apart from their psychometric appeal or inelegance, such models may be adopted or rejected by the real-life practitioners.

There is one final point about the selection models that does not seem to have received the amount of attention and consideration it deserves relative to the potentially great impact it is likely to have if ever made into official policy. That point is this: With the possible exception of the Einhorn and Bass model and the Cleary model of selection fairness, all of the models endorse the application of *double standards* for minority and majority groups. That is, two candidates for selection, acquiring the identical score on the prediction measures, will be treated differently depending upon ethnic identity. The debate among proponents of the models is largely one of how to arrive at the difference between the two standards for the two groups. Usually this difference is in the direction of lessening the requirements for the minority group, but in a fascinating demonstration of the extremes to which strictly quantitatively oriented values can lead one, McNemar (1975) has suggested that higher requirements for the minority group would be appropriate, to eliminate the overprediction that often occurs. That sort of outcome is, I'm certain, quite unacceptable to most of us, but many others will object just as strenuously to a double standard that favors minority groups, violating as it does the treasured principle of equal opportunity. Let us not, in our concern with the details and relative merits of these various new models, lose sight of the fact that a formal and official policy of double standards of any degree is likely to be met with resistance.

Test Bias as the Wrong Criterion

Gulliksen (1976) has pointed out the relation of the "criterion problem" to test bias and urged that it be given far more attention than it has in the past. It represents another instance of the theme of this article, the great complexity of the issue of test bias. The criterion problem cuts right across many of the other aspects of the problem discussed here, in particular those of differential validity and the various selection models, suggesting an awesome number of potential interactive effects.

No one really believes in the superiority of the traditional college's freshman-year grade point average or in the inviolability of supervisory ratings of job performance. Yet their presence is everywhere in these studies, like an unwelcome guest being ignored with enthusiasm.

Such criterion problems impose themselves on test bias studies in a variety of ways. On a very immediate level, the typical differences in reliability between the predictor measures and the usual criterion measures frequently cause the mean dif-

ferences between minority and majority groups to be greater on the predictor than on the criterion. This can be interpreted as bias in the predictor if no attempt is made to determine whether the mean differences would in fact be equivalent under conditions of equivalent reliability.

There is another way the criterion problem imposes itself. If we accept those traditional criterion measures for our research, attempting to maximize the correlation of an academic prediction test with these criteria operates to make the test a reflection of the traditional cultural components, at a time when such components are suffering considerable loss of prestige. This leads easily to accusations of bias, since no claim can be made for having tried to relate to innovative, nontraditional, or pluralistic educational endeavors.

Gulliksen suggests that we systematically examine the more popular criterion measures by relating them to a carefully selected battery of well-known tests. By studying the relationships, both high and low, we will be able to decide whether the criteria are really measuring the things we would like them to. On another front, some progress has been made on the question of whether the criterion measure is equivalent—measuring the same thing—for two or more groups (Rock, Werts, & Flaugher, in press), so perhaps the criterion problem is at last going to get more of the attention it deserves.

On a more abstract level, however, problems with the criterion can be seen as the reason for the legal disputes that have given the issue of test bias so much public attention recently. Some schools have adopted as a criterion of their own success the number of members of minority groups they graduate. In a sense this is a decision to modify the usual criterion of "satisfactory grade point averages" with the requirement that at least some of them be accomplished by members of a minority group. Obviously, then, this criterion cannot be met by the nonminority. The legal dispute concerns the right of the school to modify its own standards in that fashion, essentially elaborating on the more traditional criterion.

Beyond that, however, lies the question of the soundness of the criterion even as it is modified—it seems unlikely that "more minority lawyers and doctors" is desirable as a goal in and of itself; rather, it is assumed that by achieving this there will be a beneficial impact on other social goals, such as better professional services for those groups. But "more minority doctors and lawyers" is once again an approximation to that, just as we now depend on the grade point average in our studies. So, evidently, the criterion problem will always be with us.

Test Bias as Atmosphere

If persons taking a test feel out of place or unwelcome, it is reasonable to expect less than optimal test performance from them in that setting. Sex and ethnic differences can, of course, create such atmosphere bias, and this possibility definitely belongs in this list of the aspects of test bias. It may be that the very act of testing itself, whether in a hostile atmosphere or not, is unfair for some persons, in that their real capacities are inhibited when confronted by a test, and it seems reasonable that this could happen more frequently for minorities. For some reason, however, not a great deal of research has been done on this recently, perhaps because it is essentially accepted that care should be taken to see that test takers are not placed at a disadvantage because of such factors. Edgar Epps was associated with much of the earlier work in this field; he now seems convinced that beyond those basic precautions, continued study of this aspect of bias is still another one that does not promise to be productive of noteworthy results (Epps, 1974, p. 49; Epps, Note 4).

It has recently become apparent, however, that there is a very important atmosphere-type factor that may be at the base of much criticism from educators, who are certainly not the least vocal of the critics. This factor concerns the application of nationally normed examinations to certain areas of the American educational system that are in fact not in the same population, academically speaking, as the rest of the system. There are parts of the United States, primarily within the large inner-city school systems, where the majority of students are from impoverished homes, are frequently nutritionally deficient, and exist in an atmosphere of pervasive hopelessness. Within those school systems are thousands of teachers who are dedicating themselves to accomplishing whatever positive things they can in that setting. What they—students and teachers—need is support and encouragement to continue their efforts. However, sometimes the officially mandated testing provides just the opposite, serving instead to inflict on them periodic, detailed documentation of just how very far away from anything approaching the norm they really are. To witness such a painful process in operation is to conclude that something very unfair is

happening; students and teachers in that setting know this kind of test bias in a very personal and painful way and understandably are hostile, ready to condemn that process and testing itself as a demonstrably harmful influence in their lives. This then is, in a sense, another side to the concept of atmosphere bias; rather than the sort in which the test results themselves are influenced by a hostile atmosphere of discouragement and despair, this involves the creation of discouragement and despair by the process of testing itself.

It seems possible that some of the popular appeal of criterion-referenced testing as an alternative to norm-referenced testing is related to the belief that it would reduce the painful and discouraging comparisons, while maintaining the legitimate and positive function of documenting what has been learned and what standards have been met. This belief is seen in definitions of criterion-referenced tests as "achievement tests which measure exactly what has been taught" (Brazziel, Note 5) in contrast to norm-referenced achievement tests which, presumably, do not.

Although it hasn't been widely acknowledged as such, it seems that the developments in the field of tailored testing promise to provide an effective solution to this particular aspect of test bias. By using the test taker's prior performance as a basis for selecting the subsequent items, the same necessary information can be collected, the essential function of educational accountability will not be jeopardized, and yet the psychological trauma of the process can be reduced. There appears to be much more than increased efficiency at stake in the progress of tailored testing techniques.

Conclusion

It is appropriate now to review the various components of the definition of test bias that have been discussed here. First, the distinction between a test score as an indication of past achievement, rather than as an indication of aptitude for future achievement, has been seen to be an extremely important one; but this distinction is not often made in the public mind and leads to expectations that an unbiased test, for example, should show no mean differences between groups. Further, other charges of test bias originate in the tendency to overinterpret what is being measured and in the demand that tests stop reflecting the sexism that permeates our language.

Two components of the issue have received considerable research attention: differential validity, and questions about the particular content of the tests. While they are certainly legitimate aspects of the overall issue of test bias, the research results have been disappointing and indicate that these components are not as significant as some supposed.

Research on other, quite distinct components is progressing currently, and we find that those components ultimately lead us back to more careful consideration of our value systems: These are the questions of bias in the selection model and bias in the criterion measures we adopt. Finally, another sort of bias is inappropriateness, and tailored testing promises to have an impact on this.

In conclusion, it seems essential to keep in mind all of these various aspects of the definition of test bias. We continually run the risk of losing perspective on our research when we settle on one operational definition of test bias as a scientific starting point and then proceed to forget that it is only that. No matter what definition we use, because the concept is a public one we are never going to encompass all that it contains. The research on these various aspects is some of the more exciting and significant being done in psychology today, but let us not be confused about whether one of these aspects is the *real* issue of test bias. As it happens, they are *all* the real issue.

REFERENCE NOTES

1. Ebel, R. L. *Constructing unbiased achievement tests.* Paper presented at the National Institute of Education Conference on Test Bias, December 2–4, 1975, Annapolis, Maryland.
2. Ebel, R. L. *In defense of standardized testing.* Paper presented at the Houghton-Mifflin Annual Measurement Conference, March 31, 1976, Iowa City, Iowa.
3. Bianchini, J. C. *Achievement tests and differentiated norms.* Paper presented at the U.S. Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, May 27–30, 1976, Reston, Virginia.
4. Epps, E. G. *Test administration.* Paper presented at the U.S. Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, May 27–30, 1976, Reston, Virginia.
5. Brazziel, W. F. *School testing and minority children.* Paper presented at the National Education Association Conference on Test Bias, February 18–20, 1972, Washington, D.C.

REFERENCES

American Psychological Association. Guidelines for nonsexist language in APA journals. *American Psychologist,* 1977, *32,* 487–494.

Bartlett, C. J., Bobko, P., & Pine, S. M. Single-group validity: Fallacy of the facts? *Journal of Applied Psychology,* 1977, *62,* 155–157.

Boehm, V. R. Negro–white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology*, 1972, *56*, 33–39.

Boehm, V. R. Differential prediction: A methodological artifact? *Journal of Applied Psychology*, 1977, *62*, 146–154.

Breland, H. M., & Ironson, G. H. DeFunis reconsidered: A comparative analysis of alternative admissions strategies. *Journal of Educational Measurement*, 1976, *13*, 89–99.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. *An investigation of sources of bias in the prediction of job performance: A six-year study* (PR 73-37). Princeton, N.J.: Educational Testing Service, 1973.

Cleary, T. A., Humphreys, L., Kendrick, S. A., & Wesman, A. Educational use of tests with disadvantaged students. *American Psychologist*, 1975, *30*, 15–41.

Epps, E. G. Situational effects in testing. In L. P. Miller (Ed.), *The testing of black students*. Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Flaugher, R. L. *Bias in testing: A review and discussion* (TM Report 36). Princeton, N.J.: ERIC Clearinghouse on Tests, Measurements, and Evaluation, 1974.

Flaugher, R. L., & Schrader, W. B. *Eliminating differentially difficult items as an approach to test bias* (RB-78-4). Princeton, N.J.: Educational Testing Service, 1978.

Guion, R. M. Sources of bias in the prediction of job performance: Implications for governmental regulatory agencies. In L. A. Crooks (Ed.), *An investigation of sources of bias in the prediction of job performance: A six-year study* (*Proceedings of Invitational Conference*). Princeton, N.J.: Educational Testing Service, 1972.

Gulliksen, H. *When high validity may indicate a faulty criterion* (RM 76-10). Princeton, N.J.: Educational Testing Service, 1976.

Humphreys, L. G. Statistical definitions of test validity for minority groups. *Journal of Applied Psychology*, 1973, *58*, 1–4.

Katzell, R. A., & Dyer, F. J. Differential validity revived. *Journal of Applied Psychology*, 1977, *62*, 137–145.

McNemar, Q. On so-called test bias. *American Psychologist*, 1975, *30*, 848–851.

Nunnally, J. C. *Tests and measurements, assessment and prediction*. New York: McGraw-Hill, 1959.

O'Conner, E. J., Wexley, K. N., & Alexander, R. A. Single group validity: Fact or fallacy? *Journal of Applied Psychology*, 1975, *60*, 352–355.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 1976, *13*, 3–29.

Rock, D. A., Werts, C. E., & Flaugher, R. L. The use of analysis of covariance structures for comparing the psychometric properties of multiple variables. *Multivariate Behavioral Research*, in press.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests. *Journal of Applied Psychology*, 1973, *58*, 5–9.

Stanley, J. C. Predicting college success of the educationally disadvantaged. *Science*, 1971, *171*, 640–647.

Wallace, S. R. Sources of bias in the prediction of job performance: Implications for future research. In L. A. Crooks (Ed.), *An investigation of sources of bias in the prediction of job performance: A six-year study* (*Proceedings of Invitational Conference*). Princeton, N.J.: Educational Testing Service, 1972.

August 1976

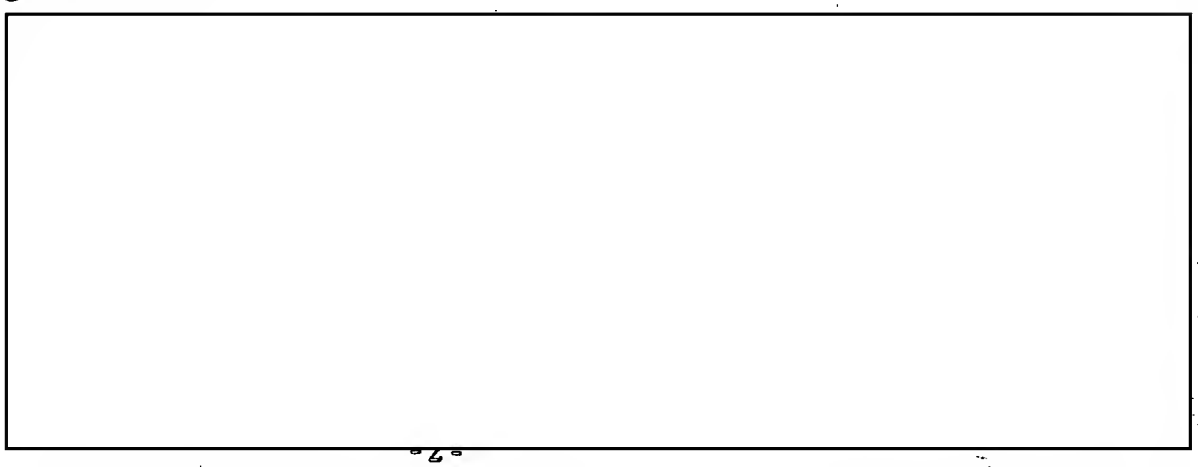
The Professional Test Battery and the Question of Fairness

This paper presents a brief history of the Agency's experience with the use of the Professional Test Battery and discusses some of the issues involved in answering the question: "is it fair?"

1. History. The Agency's Professional Test Battery (PTB*) was developed in the early 1950's as a supplement to OSS-based psychological assessment procedures for the selection and assignment of personnel. Basically, the idea was to devise a set of simply-administered group procedures which would elicit in an objective and systematic way some information relevant to an individual's abilities, aptitudes, interests, attitudes, and other aspects of his personality likely to have some bearing on his suitability for an Agency job. Tests and questionnaires were selected or developed to

*Refers to both the Professional Applicant Test Battery (PATB) and the Professional Employee Test Battery (PETB), differences between which are slight.

tap these areas, with a focus on the kinds of questions Agency officers were most interested in when they were considering a candidate for a job (How bright is he? Is he willing to work overseas in a hazardous area? Does he want to work with people? Can he write? Learn a foreign language? etc.) Gradually a package of such devices was assembled, derived from three sources: (1) standardized instruments in general use, purchasable if copyrighted or reproducible if not; (2) adaptations of standard instruments to better meet Agency needs; and (3) instruments developed solely by Agency psychologists for in-house use. The battery which emerged was tailored to a full-day's (8-hour) administration, and was used first to develop norms for Agency professional employees, such that applicants or employees taking the battery later could be compared with some kind of meaningful in-house "averages". The battery has remained fairly constant over the years, with modifications kept to a minimum in order to derive maximum advantage of the research capabilities inherent in a stable data base. The full battery consists presently of the following elements:



ST

STAT

2. Uses. It became evident early on that this battery did, in fact, yield much information that was useful in making decisions about people. Means were devised for feeding this information into the applicant selection process, and for making it available to decision-makers at other critical points in the personnel cycle. Today, information derived from all or part* of this battery is sought by many offices during the applicant consideration stage, and also at many other junctures, often when employees are being considered for a major change of duties, especially if these involve transition from one status to another (clerical to professional, contract to staff, etc.). Testing, whether for applicants or employees, is conducted at the request of the interested office, and office policy and practice vary widely. There are no Agency-wide requirements for testing as a prerequisite to selection or to any other category of personnel action.

*Applicants who are tested in the field receive only half of the battery. They may receive the second half later at Headquarters.

Test findings are typically conveyed in the form of a brief written, narrative report, descriptive in character, indicating the main highlights as they appear to relate to the pending issues to be decided. If the subject of the report is an applicant or employee under consideration for a particular assignment, the report will attempt to address considerations germane to that assignment. If the subject is a general applicant for a professional position and the report is going to the Skills Bank where any number of possibily interested offices may review it, the report may be limited to a few broad statements suggesting the general thrust of the individual's measured abilities, interests, and preferences. In no case are test scores reported as such, nor is the individual reported as "passing" or "failing". This is because there are no "cut-off scores" or arbitrarily established criteria involving test standards to which the individual must "measure up". Instead, the report presents information within the limits of what we perceive to be its basic purpose, namely, to give the decision-maker some data from a separate and unique source, to be weighed and compared with the many other types of data which he normally has at his disposal when considering a particular action. If it is consistent with other data, it may be supportive. If it is inconsistent with other data, it may trigger a useful search for the reasons behind the inconsistency. If it simply makes no sense or is not useful, it can be ignored.

3. Fairness. Does the person who takes the PTB, and thus becomes a part of this process, get a fair shake? It's a broad question. When it is asked, people tend to translate it into a question about the "validity of the tests". This is an important, but only one quite narrow aspect of the question. PSS has a rather substantial amount of data demonstrating statistical relationships between tests and what happens to people in their Agency careers. At one time or another, every single one of the components of the battery listed above have been shown to relate statistically to some measurable aspect of success in either training or job performance. In some cases, these relationships are sufficiently substantial to permit the use of a mathematical equation to yield a prediction of the odds for success in a given job. Such findings are certainly helpful, and give reassurance that we are on the right track in using tests in the way that we do. But they say very little about whether or not the system as a whole operates fairly. To tackle the broader question, it may be helpful to try to think of the many ways in which such a system could conceivably operate unfairly, and then examine the safeguards, if any, built into the system to prevent them.

4. Factors affecting test performance. In the first place, there is a basic difference in the meaning one can attach to a high performance versus a low performance on ability measures. While high scores nearly always constitute strong positive evidence of the skill or ability in question, low scores do not give

equally strong evidence of its absence. This is so because persons can score low for many different reasons. The possibilities cover quite a range, from feeling badly on the day of the test to reasons more elusive and pervasive. This means that caution in interpreting low performance is always in order. This is particularly so when the individual with low scores differs in one or more significant respects (e.g., education, socio-economic background, bi-lingual childhood home) from the normative group with whom his results are being compared. PSS psychologists can to some extent incorporate such considerations into their reporting, but must rely also on the report recipient's having some awareness of the possible impact of these factors.*

This is the general problem one faces in testing minority groups. Where minority group members test lower, one is less confident in making the same interpretations one would make of the same scores for the majority. Why not, then, stop using the tests with minorities? For two reasons: (1) Many individual minority members test well; it would be unfair to remove the positive impact of their test results from the decision process; (2) There is no good reason to believe that

*Sometimes different tests or different norms can help. For example, when increasing numbers of non-college graduates began to be referred for testing, we developed a separate version of PTB in which a different set of ability measures replaced those standardized on an Agency professional college-graduate population.

test performance/job performance relationships which hold for employees in general do not hold for minority members as well. In other words, higher test scores may predict higher job potential within both groups, even though group averages may be quite different. Data we have been able to examine thus far tend to bear this out.*

5. The question of "standards". Related also to this matter is the issue of test-related acceptance criteria which managers may impose, and the question of how reasonable or "valid" these may be. As mentioned earlier, PSS does not report test scores as such, nor describe test performance in a pass/fail mode. The recipient of the report weighs what it says and uses his own judgment in relating it to the decision he has to make. In such a system, there are no absolute barriers to prevent a given manager from using test reports as a screen to eliminate candidates who do not meet some arbitrary and unrealistic standard. Suppose, for example, that Manager A decides that he will not seriously consider any applicant for a given position who is not described in the PTB

*Legally, the issues in this area turn on (1) evidence of "adverse impact" of selection mechanisms, and (2) in cases of demonstrated adverse impact, evidence that selection mechanisms, including tests, are "valid", in the sense of measuring something clearly relevant to job performance. Court decisions have varied considerably in the stringency of their interpretations of the latter. PSS has been cognizant of EEO Guidelines in designing studies in this area in recent years.

report as less than "well above average by Agency professional norms on measures of mathematical and non-verbal abstract reasoning". If there is no convincing case to be made for the logic of such a requirement, many excellent candidates may be discriminated against unfairly. The imposition of such decisions by individual managers is difficult to guard against, or even to identify. We have no reason to believe that such practices are widespread, but we might note in general that the forces operating in organizations are such as to more often result in setting selection standards too high than too low. Rigorous examination of requirements, experimentation, and research are the only known antidotes. Affirmative Action programs sometimes spearhead such efforts. In the Agency, EEO tracking of selection processing of minority applicants is one mechanism for discouraging inappropriate uses of the PTB input.

6. User acceptance. Finally, in a program of this sort which touches many people, particularly at the applicant stage where contacts weigh heavily in one's initial impression of the Agency, the appearance as well as the substance of fairness is an important consideration. Thus, when we began to hear complaints a few years ago from female applicants who objected to a separate interest measure for women which asked seemingly irrelevant questions, we listened. Eventually, the vocational interest measures for both men and women were replaced with a single new measure for both sexes. While cost and other considerations argued against the conversion, the importance of

consumer acceptance outweighed them.

In general, reaction to the PTB by those taking it has been favorable. Applicants often comment positively on its thoroughness and breadth, and the obvious relevance of its content to legitimate Agency concerns about its prospective employees. Employees are afforded the opportunity to review the results in detail in a personal interview, and a majority of them do so. Since people change over time in all aspects measured by the PTB, re-testing is recommended frequently in preference to a re-interpretation of earlier testing which may be obsolete.

In sum, PTB, as a management tool, is subject to the same use and misuse of any other management tool. Of the thousands of applicants and employees who have taken it over the years, the vast majority have gotten a fair shake.

Prepared by:

Psychological Services Staff
Office of Medical Services

STAT

Next 6 Page(s) In Document Exempt

5. Quality control for principal products that do not contain processed information (such as bulletins of information or test books) should include inspection of a sample prior to release of the product. If the product is released from an outside vendor (e.g., outside publisher) or a sponsor's agent, quality control should include inspection of those components of the principal product that contain critical information on STAT [ ] provided services.

6. Quality control of information given in letter or telephone responses should include a periodic audit of a sample.

7. Failure to meet standards of accuracy and timeliness should be reported to a designated [ ] staff member for resolution. STAT

8. A principal product that does not meet established standards of accuracy should not be released until appropriate corrective action is taken unless release would be for the benefit of the score recipient and users and permission to release is given by STAT the cognizant [ ] officer.

9. If an error is found in critical information already released STAT by [ ] the correct information should be promptly distributed.

10. Process control methods (e.g., a predefined schedule including a delivery date and contingency procedures for dealing with volume surge) should be established for the production of each principal product to help assure its delivery by the scheduled delivery date.

11. If it is likely that there will be a substantial departure from STAT [ ] standards of timeliness with respect to a principal product, those who would be adversely affected should be so notified.

RESEARCH AND DEVELOPMENT

Principle

A continuing program of research and development conducted in compliance with professional standards with respect to quality and ethical procedures is necessary to maintain the high quality and social utility of [redacted] contributions to education. This includes basic inquiry to increase understanding of educational processes and human development; evaluative and applied research in response to the needs of the educational community; and research and development to improve [redacted] products and services. Publication of the results of significant ETS research is of benefit to [redacted] and the profession because it permits others to use, build upon or improve [redacted] work.

Policies

A. [redacted] will devote appropriate research efforts to improving education through the discovery and conceptual integration of new principles and understanding. This research will be aimed at extending knowledge of the learner and learning processes, of learning environments and educational treatments, of educational institutions and of the interacting factors that influence human development.

B. [redacted] will devote appropriate research efforts to the improvement of the technical quality of [redacted] products and services. Among the important issues addressed by this research will be problems of test development, reliability, equating, validity, and meaningfulness of interpretation.

C. [redacted] will devote appropriate research and development efforts to the identification of needs of the educational community and to the creation, improvement and evaluation of instruments, systems and programs of service that meet these needs.

STAT D. [ ] will conduct its research under appropriate procedures that protect the rights of privacy and confidentiality of human subjects or respondents.

STAT E. [ ] will follow procedures to insure that [ ] research is of high quality. Standards of quality in research refer to such matters as the identification of relevant data, the choice of suitable methods of collecting and analyzing data, the logic and objectivity of analysis and interpretation, the exploration of relationships between research problems and findings, on the one hand, and existing knowledge, theories and methodologies on the other, and the thoroughness and care of project planning and management. STAT

STAT F. [ ] will undertake research only if its potential benefits outweigh the inconveniences of or risks to the subjects or respondents who are involved.

STAT G. [ ] will encourage the dissemination of full accounts of [ ] research in the usual professional forums and will provide internal means by which the results of [ ] research can be published. STAT STAT

Procedural Guidelines

1. To maintain the quality of operational programs, [redacted] should engage in the following activities: STAT

   a) study and research on the test development process, including systematic development and evaluation of new item types and approaches;

   b) studies to determine the sources of significant differential performance of sex, ethnic, handicapped, and other relevant subgroups on [redacted] tests; STAT

   c) periodic evaluation of current approaches to aptitude and achievement measurement to determine fairness, validity and appropriateness for significant subgroups such as minorities and women;

   d) research related to reliability theory and practice, including methods of determining the reliability of classification decisions;

   e) study of the equating methods presently in use and development of improved methods as limitations in the applicability of the present methods are observed; and

   f) research to advance measurement techniques and selection and classification models relevant to fairness and validity.

2. Research projects should be undertaken in such areas as learning and cognition, personality and social influence, teacher behavior and instructional processes, socialization and human development, and the economics and sociology of education as a means of improving educational policies and practices.

3. Efforts should be made to develop instruments and programs of service in areas such as measurement, institutional

and program assessment and evaluation, instruction, guidance, financial aid, certification and licensing, and technology that would be of educational and social utility.

4. Proposals for research to be conducted by [ ] and involving human subjects or respondents should be considered by the Committee on Prior Review of Research, under its procedures for review, to verify that proper arrangements have been made for protection of the welfare and rights of human subjects. STAT

5. Researchers should not conduct research projects without the consent of subjects and respondents. In the case of young children, the consent of parents or a legal guardian, or of appropriate institutional representatives, should be obtained.

6. Each research proposal should be reviewed by one or more persons who are competent in the field within which the proposal falls. They should be satisfied that professional standards of quality and ethical conduct are met.

7. Identifiable data should be released from [ ] to researchers other than those who originally conducted the research only when one of two conditions have been met: STAT

   a) Consent to do so has been given by or on behalf of the subjects or respondents or by those who have given consent on their behalf; or

   b) the Committee on Prior Review of Research has determined that release of the data serves a public need, that there is no satisfactory and reasonable alternative way of obtaining the information, that the recipient researcher will use the data in appropriate ways and that there are adequate assurances of confidentiality.

8. After the data-collection phase of a research project has been completed, subjects should not be expected to provide additional data for a follow-up study unless such participation was part of their original agreement to serve as subjects, or their consent for follow-up is obtained or the follow-up study has been approved by the STAT Committee on Prior Review of Research.

9. The results of measures of performance based on experimental situations or tests the interpretation of which is therefore tentative and whose applied use is not yet supportable should not be reported to subjects, or to the institutions providing the subjects, unless there is relatively little danger of misinterpretation or misuse of the information that would be harmful to those individuals or institutions or unless the use is part of a feasibility study or experimental condition. Stipulations regarding nonissuance of such reports should be made to participants in advance of the data collection.

10. The results of each research project undertaken with respect to a particular STAT program or service should be available for dissemination unless a specific need to restrict publication to protect confidentiality or for other program purposes is identified prior to the beginning of the project and made known to the appropriate individuals.

11. The contracts under which research is undertaken for agencies or institutions outside STAT should permit publication of the results of the research unless a specific need to protect the research results is identified prior to the beginning of the research and made known to the appropriate individuals.

TESTS AND MEASUREMENT

This section which deals with [ ] testing activities STAT is divided into seven subsections that are devoted to test development, test administration, reliability, scale definition, equating, score interpretation, and validity.

TECHNICAL QUALITY OF TESTS

Principle

High standards of quality and fairness in constructing, administering, reporting, interpreting and evaluating [ ] tests are STAT
STAT central to [ ] capability to function effectively as an educational service and research organization.

Policies

STAT A. [ ] will strive to develop tests in which the attributes measured, procedures followed, and criteria used will be unbiased with regard to a heterogeneous group of examinees and appropriate to the use for which the test is designed.

STAT B. [ ] will establish standards for test-administration processes that minimize variations in test performance due to circumstances or conditions not relevant to the attributes being measured.

STAT C. [ ] will establish for its tests a high degree of reliability (accuracy of measurement), consistent with the requirements and the purposes of the test.

STAT D. [redacted] will develop scales for reporting scores in a rational fashion, consistent with the requirements and the purposes of the test.

STAT E. [redacted] will provide equating systems, when appropriate, for the perpetuation of scales for reporting scores at the highest level of precision practicable.

STAT F. [redacted] will make available to sponsors, institutional or agency users and examinees data for interpreting scores on [redacted] tests STAT that foster appropriate use of those scores.

G. Recognizing that test validation is a responsibility of both test users and test developers, [redacted] will encourage and assist STAT test users in their validation efforts and will itself make available tests that are designed to meet professionally acceptable standards of validity provided the use of such tests is consistent with the primary purposes for which the tests were developed.

STAT H. [redacted] will adhere to appropriate professional standards such as those published in Standards of Educational and Psychological Tests and Principles for the Validation and Use of Personnel Selection Procedures.

Procedural Guidelines

Section 1: Test Development

1. Policy and substantive contributions to the test development process should be obtained from qualified men and women who are not on the full-time staff of [redacted] and who are drawn from diverse backgrounds and appropriate specialties within professional fields (e.g., various kinds of institutions and programs, relevant philosophies and points of view, and major ethnic, handicapped and other relevant subgroups of the population). STAT

2. Appropriate background information for use in the development of a test should be documented at appropriate stages in the development process and include:

   a) the purpose for which the test is intended to be used;

   b) the nature of the population that will take the test;

   c) the relevant procedural, financial or time constraints that will influence the available test development methods and their likely outcomes;

   d) for achievement tests, the kinds of curricula for which the test is designed;

   e) for job-related tests, the elements in training or employment that are related to performance on the job.

3. For each test, specifications should be developed and reviewed by a process that provides information from the following perspectives:

a) content and skills--specifications should include the psychological, educational, or other domains to be sampled; the relative weight to be given to each domain; the appropriate level of proficiency to be required within each domain; a balance with respect to curricular differences.

b) test and item format--specifications should include the item (question) types that are most clearly related to content or skills to be measured; the appropriate level of language or reading; requirements regarding clear and comprehensive directions and sample items or the need for a sample test; and whether free-response, multiple-choice or other machine scorable formats can be used.

c) psychometric--specifications should include the level of difficulty of the test; the distribution of item difficulties (when pretested items are used); guidelines for evaluating the homogeneity among items within a test and the relationship between subtests or tests; equating requirements; number of items and time allotted.

d) sensitivity--specifications for tests should require material reflecting the cultural background and contributions of women, minorities, and other subgroups; specifications should also require a balance of positive connotations if negative connotations are made in any references to these groups.

4. Except for tests designed to measure rate of performance, the number of items in a test that has a specified time limit should be chosen so that time is not a decisive factor in performance, at least for the large majority of examinees.

5. Subject matter and measurement specialists familiar with the purpose of the test and with the characteristics of the intended population should review the test items for accuracy, content appropriateness and the adequacy with which the items sample the domain.

6. The individual items in a test should meet appropriate technical standards such as those contained in the manuals for item writers used in the test development area.

7. Individual test items and the test as a whole should be reviewed to eliminate language, symbols or content which are generally considered potentially offensive, inappropriate for major subgroups of the test-taking population or serving to perpetuate any negative attitude which may be conveyed toward these subgroups. No item in any test should include words, phrases or description that is generally regarded as biased, sexist or racist (e.g., demeaning modifiers and stereotypes).

8. The items in a test should be reviewed by editorial specialists for clarity, accuracy, consistency, and, when appropriate, for conformity with standard editorial style.

9. Tests should contain clear and complete directions. Enough sample problems should be provided in test-program publications so that the examinee can understand the nature of the task and the test-taking procedures. Where there is a need to provide a general orientation to testing, as when testing young children, practice tests--included either in descriptive material or at the time of test administration--should be used.

10. The typography, directions, and arrangement of items in the test booklet should facilitate the task of test-takers. When appropriate, tests should be made available to handicapped individuals such as sight-deficient candidates through tapes, readers or special printing.

11. Methods should be employed to evaluate the appropriateness of items before their operational use in a program or before the reporting of scores. Appropriate methods include pretesting, preliminary item analysis (using the first operational use of items as an opportunity to identify inadequate items) or careful review of the results of administering similar items to a similar population. In assessing the appropriateness of items before their operational use, efforts should be made to include representative samples of the operational test-taking population.

12. The operational use of each test should be followed by systematic item analyses using appropriate criteria and by test analyses. These analyses should include reliability, intercorrelations of sections or parts, and speededness.

13. Studies relating item performance to subgroups should be carried out for new or substantially revised tests when there are adequate data concerning sufficient samples of large subgroups whose education and experience may be different from the majority of examinees.

14. The specifications for tests in ongoing programs should be reviewed for relevance and appropriateness before each new form is created. STAT staff and advisers should consider whether changes in the field, discipline or curricula require a revision of the specifications.

15. When major changes are made in test specifications, consideration should be given to the implications of such changes for score comparability and whether it is necessary to change the test name or otherwise communicate to those who interpret test scores that comparisons with earlier tests may be inappropriate.

16. When test forms are used for a number of years in a program, they should be reviewed periodically for their appropriateness. The frequency of such review should be determined by the amount of change occurring in the population of test-takers or the subject matter domain. Test forms that are found to be outdated should be revised or withdrawn from use.

Procedural Guidelines

Section 2: Test Administration

1. Information should be made available to prospective examinees and (in some programs, to parents or guardians as well) in advance of the test administration with respect to the following, as appropriate:

   a) the purpose of the test and what it measures;

   b) the nature of the test items (including samples of typical item types);

   c) the relevant instructions for taking the test, including instructions for guessing, changing answers, and strategy involving speed and accuracy in taking the test;

   d) identification requirements and the consequences of not having identification;

   e) the consequences of misconduct by the test-taker;

   f) background and experience relevant to test performance;

   g) the location of test centers, the test dates and special testing arrangements that can be made;

   h) the procedures for registering for the test and changing the centers;

   i) the structure of test fees and fee waivers;

   j) special arrangements available for administering tests to handicapped individuals;

k) the reporting of scores;

STAT

l) procedures for canceling test scores by the candidate and reasons why [redacted] or the sponsor of the test might cancel scores; and

m) the procedures for registering complaints.

2. Program publications should be reviewed for language or descriptions generally regarded as biased and offensive. For example, the exclusive use of masculine pronouns should be avoided as should the implication that all persons in a given category (for instance, examinees, supervisors, counselors, or teachers) are either females or males (unless, of course, the category is logically restricted to members of a single sex). Illustrations, examples and practice items in test-information publications should represent males, females, minority and majority groups, and individuals in ways that indicate respect and awareness of valuable contributions.

3. The facilities at which tests are administered should be places that are convenient for the majority of examinees, nonsegregated and comfortable. At least portions of those facilities should be accessible to and responsive to the needs of handicapped individuals.

STAT

4. [redacted] should enlist test-center supervisors and staff with demonstrated sensitivity to the anticipated sex and ethnic composition of the examinee group, based on prior experience. When appropriate, persons affiliated with institutions attended by significant numbers of those examinees should be included. Minority-group supervisors and/or proctors should be employed, and test sites should be located in minority communities whenever appropriate and feasible.

5. Test-center supervisors and staff should be familiar with the procedures for administering a standardized test and should be provided with a description of the testing program, a description of the candidate population, and specific instructions for administering the test. Instructions concern such subjects as the duties of test supervisors, associate supervisors, and proctors; the receipt, storage and return of test supplies; the admittance of examinees to the testing rooms; the distribution of test materials; procedures to be followed in administering tests to handicapped individuals; procedures to be followed in instances of suspected cheating; procedures to be followed in other cases of candidate misconduct; and procedures to be followed in case of emergency.

6. Test performance can be affected by the psychological atmosphere of the testing center. Test supervisors should be informed of this and instructed to take measures to avoid an adverse situation. For example, test supervisors should be instructed, when it is appropriate and feasible, to have minority- as well as majority-group persons, women as well as men, read test directions and to recognize questions from examinees following an impartial procedure.

STAT 7. [ ] should provide the test center supervisor with directions to be read aloud to examinees before the test begins. These directions should include information relating to: procedures for marking answer sheets, timing of test sections, strategies for guessing, time and duration of test breaks and examinees' use of unauthorized aids. Test supervisors should check to see that examinees understand their task and the procedures to be followed.

8. Reasonable efforts should be made to eliminate opportunities for examinees to attain scores by fraudulent means by stipulating requirements for identification, assigning examinees to seats and requiring appropriate space between seats.

9. Appropriate procedures should be applied after the test administration to identify scores of questionable authenticity, to resolve issues of authenticity and to provide for prompt reporting of questioned scores found to be authentic.

10. A systematic program for observing test administrations should be conducted by trained [redacted] staff members or other qualified individuals: to review the testing procedures with the test supervisors, to insure appropriate testing conditions, to insure adequate maintenance of test security at the test centers and to relay questions and concerns from the field to the appropriate [redacted] office. STAT STAT

11. Testing programs should have detailed procedures for investigating and resolving examinees' complaints of irregular test administration or score reporting.

12. Comments and suggestions should be solicited from supervisors by such means as the Supervisor's Comment Sheet and meetings of supervisors to provide [redacted] staff with information to improve future administrations. STAT

13. Supervisors should be required to record and report to [redacted] information on irregularities (such as mistiming, defective materials, power failures and cheating) so that [redacted] can evaluate the possible effect of such occurrences on examinees' test performance. STAT STAT

14. An individual who has taken a test should be provided information that will be helpful in interpreting scores on that test.

Section 3: Test Reliability

1. When test scores are reported to institutional or agency users or to individual examinees, information about the reliability of the test should be documented and should include:

   a) a reliability coefficient and an overall standard error of measurement (several indices may be provided if more than one method of assessing reliability has been used; alternate-form information should always be provided if available);

   b) standard errors of measurement for score regions if decisions about individuals are made in those score regions and if the overall regions and the overall standard error are judged inappropriate;

   c) the formula(s) used to estimate reliability and/or appropriate references;

   d) a justification of the method(s) used to assess reliability;

   e) a specification of the major sources of measurement error accounted for in the reliability analysis;

   f) a specification of the time interval between testings if alternate-form or test-retest reliability is used;

   g) the number of observations, the mean and standard deviation of the analysis sample (ranges or averages are acceptable in cases where the reliability information is derived from several samples);

   h) speededness data; and

   i) correlations of subscores within the same test or battery of which the test is a part.

2. If reporting any of the reliability information required under Guideline 1 is inappropriate, the reasons should be stated in appropriate program documents and, if possible, alternate information about consistency should be provided.

3. Efforts should be made to provide reliability information in an appropriate form to the examinees to whom the scores are reported.

4. The method(s) used for assessing reliability should:

   a) take into account the most common sources of error generally considered significant for test interpretation (e.g., guessing, instability over time, item and content variation, and rater inconsistency): and

   b) be appropriate to the nature of the test, in order not to seriously over- or underestimate reliability.

Procedural Guidelines

Section 4: Scale Definition

1. Raw scores on a test or subtest (including percentages of questions answered correctly) should not be reported by [redacted] for individual examinees or in summary form for groups of examinees except under either of the following circumstances: STAT

   a) when it is anticipated that only one edition of the test will be offered for use in the foreseeable future or it is demonstrated by appropriate empirical procedures that raw scores on all the editions to be compared are interchangeable; or when raw scores on that test edition will not be compared directly with raw scores on another test edition; or

   b) when reported in conjunction with a scaled score and in a context that supports appropriate interpretation, such as when a copy of the test itself is available or when individual or group responses to individual items, depending on whether individual or group performance is being assessed, are available.

2. If a test or test battery yields multiple scores for an individual and scaled scores are to be used directly (i.e., without reference to norms tables) in interpreting performance profiles, the scales should be normatively defined and each should be defined with respect to the same population.

3. When different tests in a program are taken by different examinees whose scores are to be directly compared, the scales for the tests should take into account possible differences among the groups of examinees who take the various tests.

4. Established scales should not be redefined except under compelling circumstances. If a scale is to be substantially redefined, the numerical values should be changed substantially to minimize the possibility of confusion between test results expressed on the revised scale and results expressed on the original scale. An exception to this guideline may appropriately occur if the test in question is one of a set of tests for which a single range of numerical values (e.g., 20-80) is used and the scales for other tests in the set have not been redefined.

5. Scale properties that affect score interpretation and use should be described in program publications available to the examinees and to institutional or agency users.

6. Technical manuals and interpretive publications for institutional or agency score users and examinees should indicate, in language appropriate to the audience, whether a distributively based scale is intended to be normative or nonnormative. If it is intended to be normative, the group should be described.

7. Whenever a normatively defined scale no longer conveys useful normative information, all published descriptions of the scale should be changed accordingly.

8. Program publications should caution score recipients (users and examinees) that scores received on different tests that are reported on scales that are similar in appearance may not be equivalent.

Guidelines 9 through 14 apply only to scales established after guidelines 1-8 were published on August 1, 1977.

9. If a scale is to be distributive, the choice between a normative and nonnormative distributive scale should take into account:

   a) the extent to which normative interpretation with reference to a particular population will be appropriate and useful for all examinees who take the test and for all purposes for which the scores are intended to be used;

   b) the probable time period during which the normative information conveyed by the scores will continue to be descriptively appropriate; and

   c) the feasibility of identifying and testing a suitable group of examinees on which to base a normative scale.

10. The choice between a distributive and nondistributive scale should take into account the use for which the test was intended and to which the test is likely to be put.

11. If a scale is to be defined with reference to standards of performance, the basis for establishing the standards should be determined empirically or rationally rather than arbitrarily.

12. The conventional grade- or age-equivalent score (the grade or age for which a particular score is the average) should not be used to establish the score scale for a test or system of tests. This type of score, as it typically has been derived, should be avoided altogether as a basis for reporting test performance. However, the grade (or age) for which a particular scaled score on a test is the average, referred to here as a "grade (age) level indicator" to distinguish it from the conventional grade-equivalent (age-equivalent) score, may be reported to help in score interpretation, if the practices customarily followed in deriving and presenting grade-equivalent (age-equivalent) scores are modified in accordance with [redacted] criteria that obviate the technical interpretive problems that grade-equivalent (age-equivalent) scores create. STAT

13. The choice of a scale should take into account the likelihood of confusion with other widely used scales.

14. In establishing the number of distinct scale values to be reported, consideration should be given to the relative importance of the need to avoid erroneous distinctions among individuals (by reporting different scores for individuals whose true scores are the same) and the need to maintain distinctions that, on the average, will be correct (by reporting different scores for groups of individuals whose average true scores are different).

Procedural Guidelines

Section 5: Equating

1. Adequate equating should precede comparisons of the test performance of two or more individuals or groups on nonidentical items or sets of items such as test offerings in which successive, or alternate, forms are used interchangeably.

2. Statistical methods selected for equating should be used only under circumstances that are consistent with the assumptions under which the methods have been developed.

3. In regular and continuing testing programs that are available to users, integrated, long-range systems of equating the scores to all successive editions of the test should be used and described in technical publications.

4. For those tests that are offered for institutional use (as distinguished from externally administered tests offered in testing programs) of which only a limited number of forms are available, equating of new forms should be based on specially designed studies in which examinees or groups of examinees are selected by an appropriate sampling procedure to take the alternate forms or alternate sequences of forms.

5. When test forms are equated with the use of common (anchor) items, the psychological task of taking those items (represented, for example, by the directions, the context of the items and the speededness of the part of the test in which the items appear) should be the same for all examinees.

6. When the common items used for equating are not representative of the tests being equated, the groups of examinees used for equating should be as nearly as possible equivalent.

7. In the continuing testing programs, statistical checks (e.g., check equating, special scale-stability studies) should be employed to permit regular assessment of the precision of the equating.

Procedural Guidelines

Section 6: Score Interpretation

1. Effective test use and meaningful score interpretation should be supported and augmented by:

   a) the development of appropriate test norms based on administering tests to samples from a defined population when there is a reasonable expectation that a large proportion of the schools or other units selected for the norms sample will agree to participate; or,

   b) a rationally developed system of interpretation shared with score recipients when score interpretation is not developed from normative data.

2. Tests offered for sale and described by [ ] as standardized tests (as distinguished from tests offered in testing programs) should have adequate norms or other information for use in interpreting test results. STAT

3. When test norms are developed by administering tests to samples from a defined population, the resulting norms should be representative of any relevant subgroup, including those defined by sex or ethnicity, in proportion to their frequency in the defined population. Such subgroups may be deliberately over-sampled for more precise estimation of the statistical characteristics of the population by procedures that take over-sampling into account. Data on the proportions in the sample and in the population, when available, should be reported in an appropriate technical publication.

4. The report of a special norms study should provide information on:

   a) the sampling design;

   b) the participation rate of institutions or individual respondents in the sample;

   c) characteristics of the participating institutions and individuals;

   d) weighting systems used in preparing norms; and

   e) estimates of sampling variability along with an acknowledgment, when necessary, that such estimates do not take into account biases arising from nonparticipation.

5. When descriptive statistics based on program testing (as distinguished from norms based on special norms studies) are published, the following guidelines should be used:

   a) both table titles and descriptive material should make it clear that the statistics are based on examinees or participating institutions or other using agencies;

   b) the descriptive material should define the nature of the group by identifying the appropriateness of the sample and the factors that relate the background of the group to test performance, and by acknowledging explicitly that the sample is self-selected;

   c) when possible, reports should be prepared to show comparisons of data based on program examinees or institutional characteristics with relevant data on variables from other sources;

d) when information about interpretive data is prepared for different user groups, the presentation, whenever practicable, should be adapted to the needs and background of each group.

6. When norms are developed from program testing, the age, sex and ethnic composition of the program norms group should be described whenever such information about subgroup membership is available.

7. In testing programs, descriptive statistics should be compiled periodically from a sample or entire population in order to monitor the participation and performance of males and females drawn from diverse backgrounds, interests and experience (e.g., major ethnic group, handicapped status and other relevant subgroups of the population of interest).

8. If norms intended for use in the interpretation of individual scores are presented separately for males and females or for members of specific ethnic groups, the rationale should be carefully described. Separate norms may be justified for scores used primarily for guidance when access to the experiences needed to earn a high score is clearly related to subgroup membership and a more direct index of access is not available. The existence of score differences between subgroups does not in itself justify presentation of separate norms.

9. Descriptive statistics prepared separately for subgroups of the relevant test-taking population but not intended for use in interpreting individual scores should not be presented in a way that encourages their use for such a purpose.

10. Institutional or agency users and examinees should be informed of the standard error of measurement of a score, and test interpretation materials should point out the limitations of test scores and encourage score users to take into account the possible scores a test taker might achieve on retesting.

11. Statistical data used in score interpretation should be revised annually except when less frequent revision is judged to be appropriate as, for example, when norms are based on special studies. A statement of the period in which the data were collected should be included in any publication that presents the data.

13. Institutional or agency score recipients should be provided with interpretive materials designed to be helpful for using scores in conjunction with other information, setting cutting scores where appropriate, interpreting the scores for special subgroups (e.g., ethnic minorities, males, females, and handicapped students), conducting local normative studies, and developing local interpretative materials.

Procedural Guidelines

Section 7: Test Validity

STAT 1. [redacted] should provide evidence of the validity of its tests in relation to the principal purposes or intended uses of the tests. One or more of the following may be applicable:

a) when test scores are to be interpreted in terms of degree of mastery of the knowledge, skills, or abilities of a domain represented by the test, content validation evidence should be provided.

b) when test scores are to be interpreted in terms of the prediction of future behavior, criterion-related validation evidence should be provided.

c) when test scores are to be interpreted as a measure of a theoretical construct, construct validation evidence should be provided.

2. Evidence of content validity should be based (a) on a careful determination and analysis of the domain(s) of interest and of the relative importance of topics within the domain, and (b) on a demonstration that the test is an appropriate sample of the knowledge or behavior in the domain(s). A report on evidence of content validity should present descriptions of the procedures employed in the study, including the number and qualifications of experts involved in the analysis of the domain or evaluation of the relevance and appropriateness of the test.

3. Construct validation should be based on: rational and empirical analyses of processes underlying performance on the test in question including, where appropriate, noncognitive as well as cognitive functions. Empirical evidence relevant to the analyses should include results of investigations of the degree to which test scores are related or unrelated to other variables in ways implied by intended interpretations.

4. Criterion-related validation should be used only when technically sound and relevant criteria are available or can be developed and when other conditions affecting feasibility warrant the study.

   a) Criterion-related validation should involve as many performance variables as necessary to permit evaluation of the effectiveness of test scores for predicting the types of behavior they are intended to measure.

   b) Criterion-related validation should not combine variables to form a single criterion measure unless such a procedure is justified by logical considerations or empirical evidence or the practical requirements of the intended use of the results.

   c) Criterion data should be collected in a way that permits an assessment of the reliability of each criterion variable, but with the understanding that there may be several sources of irrelevant variation, (sampling of criterion content, source of criterion ratings or data, and so forth).

5. Interpretations of correlations between test scores and criterion variables should take into account such factors as sample size, criterion reliability, possible restriction in the range of scores obtained in the validity study sample, and other contextual factors.

6. The method(s) by which any validation is accomplished should be fully documented; such documentation should include appropriate details such as the nature and reliability of the criteria, a description of the subjects used, the materials surveyed and the qualifications of the experts who made judgments regarding the appropriateness and importance of test content.

7. Where adequate methods are employed to insure equivalence of scores on alternate forms, it is not necessary that each new form be validated. New validation studies should be made if revised tests have substantial changes, such as different item types, or if they sample a revised performance domain.

8. When appropriate and feasible, the validity of a test should be investigated separately for subsamples of the test-taking population.

9. When a name of a test is established, it should not imply more than is justified by evidence of validity.

10. Information should be made available to institutional and agency users that would be of assistance to them in planning and conducting local validity studies.

TEST USE

Principle

Proper and fair use of [redacted] tests is essential to the social utility and professional acceptance of [redacted] work.

STAT STAT

Policies

A. [redacted] will set forth clearly to sponsors, institutional or agency users, and examinees the principles of proper use of tests and interpretation of test results.

STAT

B. [redacted] will establish procedures by which fair and appropriate test use can be promoted and misuse can be discouraged or eliminated.

STAT

Procedural Guidelines

1. Program publications should:

   a) describe appropriate usès and caution against potential misuses of program tests;

   b) explain clearly that test scores reflect past opportunity to learn and discourage test interpretations that go beyond reasonable inferences from test performance;

   c) emphasize that an individual's test score should be interpreted in the context of other information about him or her;

   d) provide appropriate information about test content, difficulty, and purpose to help the institutional or agency user select instruments that meet the measurement requirements of the situation and avoid selecting, requiring or using inappropriate tests;

e) invite institutional or agency users to consult with the program sponsor and/or [redacted] about their current or intended uses of [redacted] developed tests and identify the offices to be contacted for this purpose;

STAT STAT

f) summarize results of research relevant to the use of the test or cite references in which such results are reported;

g) describe adequately and clearly scale properties that affect score interpretation and use;

h) advise institutional or agency users that decisions about the application of single or multiple prediction equations, based on distinguishing characteristics such as sex, ethnic group or curricular emphasis or training, should be preceded by careful examination of social, educational and psychometric factors;

i) advise institutional or agency users that if examinee grouping based on test scores is practiced, provision should be made for frequent review of group assignments to determine actual performance;

j) stress that pass-fail or cut-off scores established for such purposes as admission, credit, or certification, should be used as a basis for decision making only if the institutional or agency user has a carefully developed rationale, justification, or explanation of the cutting score that is adopted; and

k) encourage institutional or agency users to reexamine cut-off score policies periodically to minimize or eliminate possible disproportionate exclusion of members of any group such as men and women drawn from diverse backgrounds (e.g., major ethnic, handicapped and other subgroups of the population of interest) in the face of other evidence that would predict their success or indicate their competence.

2. Special (nonprogram) publications should be developed and disseminated by [redacted] to promote fair use of tests and discourage misuse of tests.

STAT

3. Complaints or information about questionable interpretation or use of reported scores should be investigated by means of procedures designed for detecting misuse. Such procedures should be documented, and records should be kept of such complaints and their disposition.

4. In cases where a clear misuse is brought to its attention, [redacted] should inform the sponsor and the institutional or agency user of [redacted] opinion as to the misuse and seek voluntary correction of the misuse. If reasonable efforts to seek voluntary correction are not successful, [redacted] in conjunction with the sponsor, should take steps to determine whether to continue supplying tests or reporting scores to the institutional or agency user.

STAT
STAT
STAT

TECHNICAL ASSISTANCE, ADVICE, AND INSTRUCTION

Principle

STAT [ ] is dedicated not only to providing measurement programs and conducting research but also to promoting increased understanding of measurement and test use.

Policies

STAT A. [ ] will develop and offer instructional programs in the areas of measurement, evaluation, and related research through such forms as publications, seminars, in-service training, intensive residence courses, workshops, internships and conferences. [ ] STAT may undertake these activities independently or in cooperation with other agencies, professional groups or educational institutions.

STAT B. [ ] will provide advice and information on measurement-related
STAT issues and about [ ] programs, research and services. In this
STAT activity, [ ] will work, where feasible, in collaboration with other professional organizations that show a concern about measurement.

STAT C. [ ] will respond promptly to requests for advice, instruction and technical assistance related both to programs and services
STAT offered by [ ] and to the related areas of educational measurement, evaluation and research.

STAT D. [ ] will conform to high standards of accuracy and professionalism in its advisory, instructional and technical assistance activities.

E. [redacted] will provide advice, instruction and technical assistance to clients from the private and public sectors and from foreign and domestic government agencies to the extent that such services are consistent with [redacted] areas of expertise, meet accepted professional and ethical standards, and reflect an understanding of and respect for cultural differences.

F. [redacted] will endeavor to promote increased understanding of the purposes and procedures of testing among professional groups and in the public sector; [redacted] will make this effort both independently and in cooperation with other organizations that share this responsibility.

Procedural Guidelines

1. [redacted] offices should offer advice, instruction and technical assistance; the staffing for such services should be determined by the nature of the services and the expertise required.

2. The special requirements of audiences with varying needs, interests, cultural backgrounds and levels of knowledge should be considered when [redacted] provides technical assistance, advice, or instruction.

3. New developments in research or testing should be considered when technical assistance, advice and instruction are offered.

4. Technical assistance, advice and instruction offered to institutions or agencies should include guidance on how to use other information about examinees (such as previous academic performance, English as a second language, and family or cultural background factors) in conjunction with test scores.

5. Comprehensive collections of reference materials relating to tests, measurement, evaluation and related research should be developed, maintained and made available to all [redacted] staff members and, when appropriate, to professional groups and individuals outside the organization.

STAT

GLOSSARY OF TERMS

Accuracy: The extent to which a principal product conforms to its specifications or correctly reflects the source data within the specified limits of reliability.

Client: (See Sponsor)

Consent: Permission granted by an individual or that individual's parent or guardian to the use or release of data held by [redacted] such permission granted upon receipt of a reasonable explanation of the purpose of the use or release and a reasonable explanation of the manner in which the results will be reported.

Critical Information: Information that will be used to draw important inferences (a) about the sponsor, [redacted] appointed external committees, institutional or agency user, examinee, subject or respondent, or (b) by the sponsor, institutional or agency user, examinee, subject or respondent and which, if incorrect, could be harmful.

Distributive Scale: A scale that is defined to yield either a specified score distribution or a specified mean and standard deviation for a particular group of examinees.

[redacted] Board of Trustees: The [redacted] Board of Trustees is the governing body of [redacted]. There are 16 trustees. Thirteen are elected for four-year terms. New members of the Board are elected by current trustees. Some are chosen from nominees proposed by the American Council on Education and the College Entrance Examination Board, two of the founding organizations of [redacted] The presidents of the American Council on Education, the College Entrance Examination Board and [redacted] also serve as trustees.

[redacted] held Program Data Files: Information about individuals and institutions held by [redacted] and derived from [redacted]-provided services of collection, processing, storage, retrieval and dissemination.

[redacted] held Research Files: Information held by [redacted] and generated through [redacted]-conducted research intended to result in the development of new or improved techniques and materials for application in such areas as classroom instruction, evaluation of progress toward educational goals, counseling of students, and decision-making of school administrators.

Examinee: An individual who takes a test, developed and or administered by [redacted]

Institutional or Agency User: An organizational recipient of [redacted] processed or produced information.

Intermediate Product: Materials that are not released externally, but that are necessary to the production of the principal product.

Glossary (continued)

Nondistributive Scale: A scale that is defined without reference to the observed test performance of a particular group.

Nonnormative Scale: A scale that is based on the performance of any conveniently available subgroups of examinees for whom the test is appropriate. A score on a nonnormative scale is not intended to convey information about an examinee's standing in relation to a defined population.

Normative Scale: A scale that is based on the test performance of a sample of examinees, selected as prescribed by a specified design, from a clearly defined population. A score on a normative scale is intended to convey useful information about the performance of a particular examinee in relation to the performance of that population.

STAT Principal Product: [redacted]-produced or processed materials (e.g., annual reports, performance data, score reports and admissions tickets) that are released or transmitted to a sponsor, [redacted]appointed external committee, institutional or agency user, examinee, subject or respondent, pursuant to a contract or published commitment. Standards with respect to accuracy and timeliness are applicable to principal products. STAT

Principles for the Validation and Use of Personnel Selection Procedures, Division of Industrial-Organizational Psychology, American Psychological Association. Dayton, Ohio: The Industrial-Organizational Psychologist, 1975.

Respondent: An individual who provides data to a research project in a manner and for a purpose different from either examinees or subjects.

Sponsor: Educational, professional or occupational associations, federal, state or local agencies, public or private foundations
STAT which contract with[redacted]for its services. This category includes their governing boards, membership, and appointed committees or staff.

Standards for Educational and Psychological Tests, American Psychological Association (APA), American Educational Research Association, and National Council on Measurement in Education. Washington, D.C.: APA, 1974.

Subgroup: A part of the larger population which is definable according to various criteria as appropriate, e.g., by (a) sex, (b) race or ethnic origin, (c) training or formal preparation, (d) geographic location, (e) income level, (f) handicap, (g) age.

Glossary (continued)

Subject: an individual who participates in an [ ] laboratory or experimental research project. STAT

Testing Program: A set of arrangements under which examinees are scheduled to take a test under standardized conditions, the tests are supplied with instructions for giving and taking them, and arrangements are made for scoring the tests, reporting the scores, and providing interpretative information as part of a comprehensive ongoing service. A program is characterized by its continuing character and by the inclusiveness of the services provided.

Timeliness: The degree to which a principal product is released or delivered to its recipient within a predefined schedule.

0016301 · S39P3 · 292208 · Printed in U.S.A.

Approved For Release 2004/05/12 : CIA-RDP83-00156R000300050004-6